ANALYSIS OF SCHEDULING POLICIES FOR A M/G/1 QUEUE WITH REWORK

THESIS

Jennifer K. Hendrixson, First Lieutenant, USAF

AFIT/GOR/ENS/03-09

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense or the United States Government.

AFIT/GOR/ENS/03-09

ANALYSIS OF SCHEDULING POLICIES FOR A M/G/1 QUEUE WITH REWORK

THESIS

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Operations Research

Jennifer K. Hendrixson, B.A.
First Lieutenant, USAF

March 2003

AFIT/GOR/ENS/03-09

ANALYSIS OF SCHEDULING POLICIES FOR A M/G/1 QUEUE WITH REWORK

Jennifer K. Hendrixson, B.A.
First Lieutenant, USAF

Approved:

Dr. Jeffrey P. Kharoufeh Thesis Advisor	Date
Dr. James W. Chrissis Committee Member	Date

Acknowledgments

I would like to thank the following people who have contributed significantly to the completion of this thesis: Dr. Jeff Kharoufeh, for all of his guidance, wisdom and patience, Dr. James Chrissis, for his revisions and valuable input, Capt Don Hoffman, Capt Joe Price, Capt Brad Beabout, for their assistance with ARENA and real-world USAF examples of a multi-class M/G/1 priority queue with optional rework. Additionally, I would like to thank Lt Nicholaus Yager for his valuable help with coding.

Most of all, I thank my family for their endless love and support.

Jennifer K. Hendrixson
March 2003

Table of Contents

List of Tables

AFIT/GOR/ENS/03-09

Abstract

This thesis analyzes a multi-class M/G/1 priority queueing system in which distinct job types require one service cycle and, with non-zero probability, require a second service cycle. The main objective is to find a new heuristic scheduling policy that minimizes the long-run expected holding and preemption costs. Arrival rates, service rates, and the probability of undertaking second service are all class specific. A mean value analysis (MVA) approach was employed to derive the long-run mean time in queue for each job type under each policy, thereby providing the appropriate cost equations. Numerical experiments suggest that the *preemptive resume* scheduling policy yields the lowest cost most frequently.

ANALYSIS OF SCHEDULING POLICIES FOR A M/G/1 QUEUE WITH REWORK

1. Introduction

1.1 Background

A queueing system can be described as any system in which arriving entities, jobs, customers and so forth, place demands on a finite-capacity resource, such as a server. Arrivals and resource demands are realistically stochastic, which may cause a queue to form. The types of entities requiring service may be either single-class or multi-class. A single-class queue has only one job type entering the system, whereas a multi-class queue has more than one job type entering the system. One example of a multi-class queue is a toll booth on a turnpike. Several types of vehicles pass through the system; the number of axles a vehicle has could be used to determine the class of the vehicle. Taking it one step further, different classes could take on different priorities. In reference to the toll booth example, if two-axle vehicles are served before five-axle vehicles, then two-axle vehicles may be said to have priority over five-axle vehicles.

Another characteristic of a queue is the number of service cycles a particular job requires. For instance, a job may require only one service cycle, or its completion may require more than one service cycle. In a manufacturing setting, a part might need to pass a particular quality test, however, if the part fails the test, the part is sent back for rework. This is referred to as a *re-entrant* queue, or a queue with *rework*.

Queues are also characterized by the queueing discipline they follow. Some examples of queueing disciplines include first-come, first-served (FCFS), last-come, first-served (LCFS), random order, and various priority disciplines. A priority discipline is one in which the jobs are assigned priorities based on some characteristic, such as cost or service time (or number of axles), upon entering the system and then served according to highest priority first. The priority scheme can be non-preemptive, wherein the higher priority job moves to the front of the line but it does not preempt a lower priority job that is already in service. If the priority scheme is preemptive, the higher priority job preempts a lower priority job already in service. The lower priority job that is preempted from service can then either repeat all necessary service time, which is referred to as *preemptive repeat*, or the service time can be resumed where it left off, which is referred to as *preemptive resume*. In this thesis, a preemption discipline is assumed to be preemptive resume, unless otherwise stated.

A queueing system with a single-server, exponential interarrival times and generally distributed service times is classified as an M/G/1 queue. This work addresses a multi-class M/G/1 queue with optional rework. More specifically, there exists a non-zero probability that a job completing its first service needs second service. A job cannot re-enter the queueing system more than once; only one rework is allowed. All arrival rates, service times and probabilities of re-entry are class specific. In addition, there are various associated costs for each job type. Specifically, there are holding costs and preemption costs. Holding cost is incurred when an entity waits in the queue, and the preemption cost is incurred when an entity in service is preempted.

A real-world example of this type of system can be seen in the USAF aircraft maintenance arena. Consider an avionics backshop, where line replaceable units (LRUs) are sent for repair from the flightline (*i.e.*, the "front shop"). The LRUs are the entity types and the server is an automatic test station (ATS). The ATS is designed to test and repair different types of LRUs from the same aircraft subsystem. For example, the ATS could be designed to repair the radar system on the

F-16 which is comprised of several different LRUs (all of which comprise the radar subsystem). There are several ways the operator of the ATS can schedule incoming LRUs depending on the attributes of the LRUs. For example, if the setup and run times on the ATS are lengthy and no priority is given to different LRUs, FCFS policy might be a reasonable choice. However, if priority is given to different LRUs, then it would be beneficial to determine which priority scheduling policy to employ to minimize overall costs.

The scheduling policy adopted for the queueing system determines the total cost, or time expended for service for any particular job type or for the the facility as a whole. Thus, the overall objective is to find the optimal scheduling policy to achieve the minimum costs or the minimum overall time in system. This thesis considers a new scheduling policy to minimize total costs in a multi-class M/G/1 priority queue with optional rework. This new policy is compared against three other scheduling policies, specifically first-come, first-served (FCFS), non-preemptive priority (NPP), and preemptive resume (PR). First, all of the relevant performance measures are derived using a mean value analysis (MVA) approach, and then the cost equations are derived and a numerical experiment is performed to determine the policy that performs the best most often. Taking an MVA approach allows the appropriate performance measures to be derived using only the first and second moments of the underlying service distributions. The MVA approach is simpler than seeking the performance measures directly via a transient analysis.

1.2 Problem Definition and Methodology

Various transportation, communication and manufacturing systems have distinct jobs flowing into a mechanism where they receive some sort of service. Upon completion of the service, with a non-zero probability the job must re-enter the system to receive a second service. This thesis models such a system and compares four different scheduling policies to determine the scheduling policy that minimizes certain costs most often. There are three different costs used in the comparison: holding costs, preemption costs when a job is preempted during service, and preemption costs when a job is preempted between first and second services. The costs to preempt during service is assumed to be much greater than the cost of preemption between first and second service. This principle can be seen in real-world situations in the manufacturing setting. Due to set-up costs, it is usually more costly to interrupt the service of a part than it is to wait until the part has completed first service. Currently, it is unknown which of the four scheduling policies minimizes costs.

The objective of this thesis is to analyze a multi-class M/G/1 priority queue in which distinct job types require one service cycle, and with non-zero probability require a second service cycle. The results of the analysis are then used to compare the total expected costs of the system under FCFS, NPP, PR, and a new (heuristic) scheduling policy. Arrivals, service times, and the probability of undertaking second service are all class specific. Only one rework is allowed. All policies are assumed to be work conserving. [1]

This research compares four scheduling policies to determine the best policy to minimize costs and examines the impact of a few factors on the minimum cost structure for a multi-class M/G/1 priority queue. Two types of costs are considered, holding costs and preemption costs. The results can possibly be used in the Air Force for scheduling different types of training, aircraft maintenance or communication systems.

The objectives of this thesis are met by first deriving the relevant performance measures for the three known policies, FCFS, NPP, and PR. Next, the relevant performance measures for the new policy, to only allow for preemption between first and second services, are derived. The cost equations for the four different policies are

[1] A work conserving policy is one in which the server works if there is a job in the system and no job reneges. In other words, there is never a queue unless the server is busy and a job cannot leave the queue, unless moving into service.

then derived. Finally, a numerical experiment is conducted to statistically compare the costs of the four different policies and to determine the best policy to adopt.

1.3 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 gives an overview of previous work and the relevant queueing literature. In Chapter 3, a mathematical model is developed and performance measures of a multi-class M/G/1 queue with optional second service are derived. In Chapter 4, the numerical experiment will be presented, along with some conjectures. Finally, Chapter 5 presents conclusions of the thesis and recommendations for future research pertaining to the scheduling of multi-class queues with re-entrant lines.

2. Review of the Literature

This thesis addresses a multi-class M/G/1 priority queue with optional rework. The system has only one server, unlimited system capacity, and infinite job population [21]. The relevant literature can be divided into two sections. The first section covers M/G/1 queueing systems without rework; the second section covers M/G/1 queueing systems with rework.

2.1 M/G/1 Queue Without Rework

The standard M/G/1 queue has been studied extensively [33]. Kleinrock's [21] text is one of several (*e.g.* [32],[31], [17]), that study classical queueing theory and derive the performance measures of a M/G/1 queue.

Kleinrock uses the method of embedded Markov chains developed by Kendall [19] to obtain steady-state performance measures. However, there are other methods used to derive performance measures, such as the method of supplementary variables [7]. For the method of supplementary variables, the state of the system is described using two variables, one that indicates how many jobs are in the system and one that indicates the expended service time of the job currently in service. For the embedded Markov chain technique, the goal is to be able to describe the state of the system with a Markov chain and only use one variable. If the system is only considered at the time of job departures, then the number of customers left behind by the departing job forms a Markov process [21]. Kleinrock [21] presents a proof that the stochastic process describing the number in system is, in fact, a Markov chain. Kleinrock uses the embedded Markov chain to derive the steady-state performance measures of the M/G/1 queue. Currently, all results for a single-class M/G/1 queue are well known.

For a multi-class M/G/1 queue, priority disciplines are often used. A multi-class queue contains more than one class of jobs, each with a distinct priority, where arrival rates and service times are class specific. Jaiswal [18] devoted an entire text to priority queues. The results presented in his book are for a variety of multi-class, single-server queueing systems. The queues studied by Jaiswal followed various priority disciplines, such as preemptive resume, preemptive repeat, FCFS, and several others. The method used in his text is primarily *busy period analysis.* This technique analyzes the evolution of the queue as a sequence of busy periods and idle periods [18]. The expected remaining service time of the job currently in service is an essential component in the long-run average wait time for a priority queue. Renewal theory has contributed to this aspect of priority M/G/1 queues [14], [13]. Fakinos [13] derived the expected remaining service time for a single-server queue using renewal theory. In his paper, he uses a result from Green's article [14] on interrenewal times to show the expected remaining service of the job currently in service when an arbitrary job arrives to the system. Renewals occur at the time of job arrivals, and the residual life is the remaining service time.

Networks of M/G/1 queues have received considerable attention in the priority queue realm (*e.g.* [2], [12], [1]). A queueing network is a directed graph in which each individual node is a queueing station, or center, and arcs exist between two nodes only if jobs may proceed between such nodes. In Baskett, *et al.* [2] the joint equilibrium distribution of queue sizes, or the average number in queue at each station, is derived. The network contains N service centers and R classes of customers. In this paper, the authors explore dynamic class assignments in a network. In the network, different queueing disciplines such as FCFS, LCFS, no queueing, and processor sharing are associated with the different service centers. The model this paper considers includes one service cycle.

Bryant *et al.* [6] use an MVA priority approximation to estimate the steady-state performance measures of a priority queueing network. In this method, the exact means of the distributions are used to approximate performance measures of a network. Eager and Lipscomb [12] later developed a more accurate and computationally efficient approximation to the priority queueing network with multiple classes. Eager and Lipscomb [12] built upon the MVA priority approximation developed by Bryant, *et al.* [6] to propose a new algorithm to obtain performance measures of a queueing network. Instead of exact means, Eager and Lipscomb [12] used approximate means to develop the approximate mean value analysis (AMVA) method to estimate the performance measures of a priority queueing network.

Groenevelt *et al.* [15] give a closed-form expression for the long-run average expected cost and the bias vector in a two-class exponential preemptive resume priority queue with holding and switching costs. The authors demonstrate the one-step policy iteration algorithm provides a near-optimal policy for the given system [15].

In the recent literature, M/G/1 priority queues with vacation times have been studied. This type of queue assumes that the server goes on vacation (or is idle) whenever the system is in a certain state. A survey of such queueing systems can be found in Doshi [11]. One extension of the vacation time method is the N-policy method. N-policy methodology assumes the server is turned off when the system is empty and turned on again when the number in system reaches N and stays on until the system is empty. Artalejo [1] discusses the M/G/1 queue with N-policy and develops a stochastic decomposition property for the waiting time. Artalejo's results hold for a single-class M/G/1 FCFS queue with N-policy. Clearly, the N-policy is not a work conserving policy since the server only starts work when there are at least N jobs in the queue.

2.2 M/G/1 Queues With Rework

Recently, much research has been devoted to M/G/1 queues with rework, particularly in the area of queueing networks.

Kumar's [23] seminal paper on queueing systems describes re-entrant lines in a manufacturing setting. Kumar [23] defines re-entrant lines to be those in which some of the jobs require rework. In this paper, Kumar proves the stability conditions and the performance measures for several scheduling disciplines for networks with re-entrant lines. There are several machines with buffers that need to be scheduled. The objective is to minimize the mean cycle time or the variance of the mean cycle time. He then provides bounds for the mean delay time and a schedule to reduce mean delay time. Kumar provides results for various scheduling policies; no direct comparisons are made, and he does not suggest an optimal policy.

Youngshin Park *et al.* [29] propose an approximation for estimating the performance measures of the re-entrant queue with single-job machines and batch machines using an MVA approach. In this paper an approximation is given for the steady-state averages of cycle time, throughput, and queue length at each station. These results are an extension of the ideas first proposed by Narahari and Khan [28] who used MVA to approximate performance measures of re-entrant manufacturing systems.

Youngshin Park, *et al.* [30] later extend their previous research [29] to include the case of re-entrant lines with multi-class jobs and multi-server workstations. In this article, a queueing network with both batch machines and single-server machines is analyzed using an MVA approach. A single-server machine only processes one job at a time, whereas a batch machine processes several jobs at a time. The authors propose an approximation method for estimating the performance measures of a multi-class, re-entrant queueing network. In this paper, the scheduling policy at each station is assumed to be FCFS.

Bard, *et al.* [4] address the problem semiconductor manufacturers face when deciding how much capacity to build into their systems, while having to meet budgetary constraints. Four new approximation algorithms were developed that take into account re-entrant flow in manufacturing systems. These algorithms are used to solve a non-linear integer mathematical program and provide approximations for the performance measures of a single-class queueing system. The algorithms have been tested on fabrication facilities that experience long cycle times, high inventories, and poor on-time performance. The decision of which algorithm to employ is based on the size of the queueing network being analyzed. For smaller networks, a particular algorithm is might be appropriate, whereas for larger networks, a combination of the four algorithms might be appropriate.

Zargar [34] studies the effect of rework strategies on cycle time. Batch manufacturing is considered in which, with some probability, a portion of the batch will be reworked. Two different strategies are discussed. The first strategy is holding back the "mother" lot to wait for the "children," and then all units are reunited and continue flowing through the network. The second strategy is to hold the "child" back, while the "mother" lot moves forward. In the second case, there are three options for handling the jobs requiring rework: the reworked parts could continue through the system alone, a reworked job could wait until some minimum order quantity of reworked units is needed, then moves on, or the reworked units could be added to the next lot and continue flowing through the system. Zargar [34] develops these policies to determine the effect on the cycle time. It was found that the policy chosen affects cycle time significantly. Moreover, it was found that adding the reworked units to the next mother lot is the optimal policy when rework probabilities are small.

Grosfeld-Nir and Gerchak [16] consider serial manufacturing facilities with repeated rework at each station. The authors propose an optimal policy to minimize the expected cost associated with each order. Costs include set-up costs and a linear variable cost. Their results show that a multistage system with only one stage requiring rework can be reduced to a single-stage station. In addition, they also propose the optimal way to arrange a multi-stage system, if given a choice, is to put the rework stage first.

Lin, *et al.* [25] explore the random inspection rate for a flexible assembly system (FAS). The authors analytically model a FAS to determine how to maximize profits and decide when to perform in-process random inspections. Their model includes a single-class network with feedback and rework at every stage. The scheduling policy adopted for their model is not discussed.

Another current trend in queueing theory is to model heavy traffic systems as fluid queues. A queueing system in heavy traffic implies that the traffic intensity is approaching unity. Dai [8] was one of the first to model a queueing system as a fluid model and used the stability condition of queues as the deterministic fluid limit. The fluid model approach was also applied by Ball, *et al.* [3] in their paper on robust feedback for a single-server with no re-entry. Day, *et al.* [9] extended the fluid model Ball, *et al.* [3] used to include re-entry. Diaz-Rivera, *et al.* [10] also consider a re-entrant manufacturing system modelled as a fluid network and prove that the long-run dynamics are periodic orbits and non-chaotic. The authors develop a method to calculate the period of the orbits, thus proving the non-chaotic behavior of a fluid network.

Most relevant to this thesis is the paper by Madan [27] who addresses a single-class M/G/1 queue with rework. Madan [27] derives the long-run expected wait time in queue for a single-class M/G/1 queue where each job has a given probability of rework. First, Madan [27] states the differential-difference equations for the system, then derives the time-dependent probability generating function of the queue length via Laplace transforms. Madan [27] then derives the steady-state results explicitly from the time-dependent probability generating functions. He validates his results by comparing them with known results for specific queueing systems. However, the results apply to only a single-class queue with some restrictions on the service time distributions.

The scheduling policy adopted for a particular priority queueing system can be significant when determining the total cost of service for any particular job type or for the the facility as a whole. Thus, the overall objective is to find the best possible scheduling policy to achieve the minimum expected costs in the long-run. This thesis considers a new scheduling policy that has not been addressed in the literature that attempts to minimize the expected total costs in a multi-class M/G/1 priority queue with optional rework. This new policy is compared against three other scheduling policies, specifically first-come, first-served (FCFS), non-preemptive priority (NPP), and preemptive resume (PR). The scheduling disciplines FCFS, NPP, and PR have been studied in the literature for multi-class M/G/1 priority queues without rework. The performance measures derived in this thesis incorporate the possibility that an entity may require a second service cycle.

It has been shown in this brief review that the multi-class M/G/1 queue without rework has been studied extensively. Additionally, networks of M/G/1 queues have been studied with and without rework. However, very little research has been performed to address the multi-class M/G/1 queue with optional second service. Although, Madan [27] addressed this type of system in his paper, the author assumes the second service is always exponential. No such assumption has been made in this work. Also, he assumes all jobs are of a single class, so there are no preemptions, and the scheduling policy is FCFS. This thesis extends the results of Madan [27] to include two classes of jobs, the second service is generally distributed, and various scheduling policies are examined. Many manufacturing, communication and transportation systems can be modelled as multi-class M/G/1 priority queues with optional second service; thus the results of this thesis can be highly applicable in a variety of settings.

In Chapter 3, the formal mathematical model is presented and performance measures derived for a multi-class M/G/1 queue with optional second service. First, the overall long run expected service times are developed for the four scheduling policies, then the performance measures and cost equations are derived for the purpose of comparing the four policies.

3. Formal Mathematical Model

The process of scheduling within a single-station queueing system involves several potential queueing disciplines such as first-come, first-served (FCFS), last-come, first-served (LCFS), random-order, batch service, non-preemptive priority (NPP), and preemptive resume priority (PR). This thesis compares the total overall expected costs of FCFS, NPP, and PR disciplines to the overall expected costs of a new scheduling policy. The new scheduling policy only allows preemption between first and second service cycles, and is developed herein. The performance measures for FCFS, non-preemptive priority and preemptive resume, along with the new policy, are derived herein. For a more in-depth discussion of the FCFS, NPP, and PR policies, the reader is referred to Gross and Harris [17], Kleinrock [22] or Jaiswal [18].

The purpose of this thesis is to compare the total long-run expected costs for the FCFS, NPP, PR and the new policy, which is then used to heuristically determine the policy that yields the minimum cost most often. FCFS is discussed first, then non-preemptive priority, next preemptive resume, and finally, a new policy is derived. Only the two class case is considered here. Throughout this thesis, a job of type i is assumed to be a higher priority than a job of type j if $i < j$.

3.1 Results for a Standard M/G/1 Queue

The following nomenclature applies to a single-station, multi-class M/G/1 queueing system. Arrivals are according to a Poisson process and the service time distribution is general. Define,

$k =$ Number of classes; $k = 1, 2...K$

$E[S_i] = 1/\mu_i =$ mean service time for a job of type i; $i = 1, 2..., k$

$E[S_i^2] =$ Second moment for a job of type i; $i = 1, 2..., k$

$B_i(\cdot) =$ CDF of service time for a job of type i; $i = 1, 2..., k$

$S_i =$ Service time for a job of type i; $i = 1, 2..., k$

$S =$ Service time of an arbitrary customer arriving to the system

$\lambda_i =$ Arrival rate for a job of type i; $i = 1, 2..., k$

$\lambda = \sum_i \lambda_i$

$\rho_i = \lambda_i/\mu_i$

$\rho = \sum_i \rho_i$

$\Lambda =$ Class type of random customer

$P\{\Lambda = j\} = \lambda_j/\lambda$

$W_i^A =$ Long run average time in system for a job of type i, $i = 1, 2..., k$ under policy A

$W_{iq}^A =$ Long-run average time in queue for a job of type i, $i = 1, 2..., k$ under policy A

$L =$ Long-run average number in system

$L_q =$ Long-run average number in queue.

The Pollaczec-Khintchine (P-K) formula for average long-run number in system is [17]

$$L = \rho + \frac{\lambda^2(\sigma^2 + \frac{1}{\mu^2})}{2(1-\rho)}. \tag{3.1}$$

For the case $k = 2$, W_q, W_1, W_2 are sought. First, L is derived, and then Little's Law is applied [26]. However, to apply Little's Law, $E[S]$ and $E[S^2]$ must be known.

First, $E[S]$ is derived by using a simple conditional expectation argument.

$$\begin{aligned} E[S] &= \sum_i E[S|\Lambda = j]P\{\Lambda = j\} \\ &= \sum \frac{E[S|\Lambda = j]\lambda_j}{\lambda} \\ &= \frac{1}{\lambda}\sum E[S_j]\lambda_j \qquad (3.2) \end{aligned}$$

where $E[S_j] = \mu_j^{-1}$. For the case $k = 2$

$$E[S] = \frac{\lambda_1 E[S_1] + \lambda_2 E[S_2]}{\lambda}. \qquad (3.3)$$

In a similar manner, it is seen that

$$\begin{aligned} E[S^2] &= \sum_i E[S^2|\Lambda = j]P\{\Lambda = j\} \\ &= \sum \frac{E[S^2|\Lambda = j]\lambda_j}{\lambda} \\ &= \frac{1}{\lambda}\sum E[S_j^2]\lambda_j. \qquad (3.4) \end{aligned}$$

Again, for the case $k = 2$

$$E[S^2] = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{\lambda}. \qquad (3.5)$$

The results for $E[S]$ and $E[S^2]$ hold for all work-conserving policies [21]. A policy is considered work-conserving if the server works as long as a job is in the system and no job reneges and leaves the system without first receiving service.

3.1.1 First Come First Served (FCFS)

The first policy (scheduling discipline) considered is first-come, first-served (FCFS). In this case, L is found using the Pollaczec-Khintchine formula, and then W_q, W_1, and W_2 can be determined.

To determine L, the famous Pollaczec-Khintchine formula is applied,

$$L = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \tag{3.6}$$

Dividing by λ yields

$$W^{FCFS} = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}, \tag{3.7}$$

which in turn yields the average waiting time in queue for an arbitrary job,

$$W_q^{FCFS} = \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{3.8}$$

Because the queue discipline is FCFS, the wait in queue is the same for jobs of class 1 and 2 and the total time in system for jobs of class 1 and 2 are, respectively,

$$W_1^{FCFS} = E[S_1] + W_q \tag{3.9}$$

and

$$W_2^{FCFS} = E[S_2] + W_q. \tag{3.10}$$

To be complete, the average number of jobs for both classes in the queue is

$$L_q = \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \tag{3.11}$$

3.1.2 Non-Preemptive Priority (NPP)

Relevant performance measures for the non-preemptive priority (NPP) discipline are reviewed. For this service discipline, whenever a higher-priority job enters the system it is moved to the front of the queue, regardless of other jobs that arrived first. However, if a lower-priority job is in service, the higher priority job does not preempt. For example, if a job of type 2 is in service when a job of type 1 enters the system, the job of type 2 completes service and then the type 1 job receives service [18]. To derive the queueing performance measures, define the following

$\mathcal{W}_i^q$ = random waiting time in queue for class i

U = remaining service time for current job in service

ψ_j = random time to serve all jobs of type j

T_j = random time to serve all jobs in time interval $[0, \mathcal{W}_i^q]$.

Assume the system is in steady-state. The time in queue for an arbitrary job is the remaining service time for the current job in service, plus the time to serve all higher-priority jobs that are in the system upon arrival, plus the time to serve all higher-priority jobs that arrive in the system while job i waits to be served. Mathematically, the time in queue is a random variable expressed as

$$\mathcal{W}_i^q = U + \sum_{j=1}^{i} \psi_j + \sum_{j=1}^{i-1} T_j \tag{3.12}$$

for $i = 1, 2...k$. Using a mean-value analysis approach,

$$E[\mathcal{W}_i^q] = E[U] + \sum_{j=1}^{i} E[\psi_j] + \sum_{j=1}^{i-1} E\left[T_j\right] \tag{3.13}$$

It is known that [17]

$$E[U] = \frac{\lambda E[S^2]}{2} \tag{3.14}$$

$$\begin{aligned} E[\psi_j] &= \lambda_j E[\mathcal{W}_j^q] E[S_j] \\ &= \rho_j E[\mathcal{W}_j^q] \end{aligned} \tag{3.15}$$

$$\begin{aligned} E[T_j] &= \lambda_j E\left[S_j\right] E\left[\mathcal{W}_i^q\right] \\ &= \rho_j E\left[\mathcal{W}_i^q\right] \end{aligned} \tag{3.16}$$

Substituting (3.14),(3.15), and (3.16), gives

$$E[\mathcal{W}_i^q] = \frac{\lambda E[S^2]}{2} + \sum_{j=1}^{i} \rho_j E[\mathcal{W}_j^q] + \sum_{j=1}^{i-1} \rho_j E[\mathcal{W}_i^q]. \tag{3.17}$$

After simplifying and solving for $E[\mathcal{W}_i^q]$, the expected waiting time in queue for a job of type i is obtained as

$$E[\mathcal{W}_i^q] = \left[1 - \sum_{j=1}^{i-1} \rho_j\right]^{-1} \left[\frac{\lambda E[S^2]}{2} + \sum_{j=1}^{i} \rho_j E[\mathcal{W}_j^q]\right]. \tag{3.18}$$

Suppose there are only two job types ($k = 2$). For case 1, $i = 1$

$$W_{1q}^{NPP} = \frac{\lambda E[S^2]}{2(1-\rho_1)}. \tag{3.19}$$

Notice, this equation takes into account the random service time of a job already in service as well as the effect of the priority of the arriving job. For case 2, $i = 2$ and

$$W_{2q}^{NPP} = \frac{\lambda E[S^2] + 2(\rho_1 E[W_{1q}^{NPP}] + \rho_2 E[W_{2q}^{NPP}])}{2(1-\rho_1)}. \tag{3.20}$$

After simplification, and applying the work-conserving principle,

$$W_{2q}^{NPP} = \frac{\lambda E[S^2]}{2(1-\rho)(1-\rho_1)}. \tag{3.21}$$

Once again, the equation takes into account the random service time of a job already in service, in addition to the effect of higher priority jobs arriving to the system. Observe, in the 2-class case, the highest priority job sees a FCFS queue, while the lower priority job's time in queue depends on both classes.

3.1.3 Preemptive Resume (PR)

For the preemptive resume discipline, if $j < i$, then class j preempts class i and completes service. Class i jobs then resume service where they left off. For example, if a class 2 job is in service when a class 1 job arrives, the class 1 job seizes the server immediately. Upon re-entering service, the class 2 job resumes service where it left off, and no service time is lost. Since preemptive resume is a work conserving policy, Little's Law [26] still holds. To derive the performance measures, define:

$L_i =$ Average long run number in system for class i

$W_i =$ Average long run time in system for class i

$W_{iq}^{PR} =$ Average long run time in queue for class i under preemptive resume policy

$L_{iq}^{PR} =$ Average long run queue length for class i under preemptive resume policy

$W_i(t) =$ Total work content in system of type i at time t.

Define the following:

$$\lambda(i) = \sum_{j=1}^{i} \lambda_j, \qquad 1 \leq i \leq k \tag{3.22}$$

$$E[S(i)] \;=\; \sum_{j=1}^{i} \frac{\lambda_j E[S_j^2]}{\lambda(i)} = \frac{1}{\mu(i)}, \tag{3.23}$$

$$\rho(i) = \frac{\lambda(i)}{\mu(i)}, \tag{3.24}$$

and

$$\rho = \sum_{i=1}^{k} \rho(i). \tag{3.25}$$

For the first i classes the total work content at time t is

$$Y_i(t) = \sum_{j=1}^{i} W_j(t) \qquad 1 \le i \le k. \tag{3.26}$$

Assume $Y_i(t) \to Y_i$ in distribution as $t \to \infty$, then using mean-value analysis (MVA), the following result is obtained

$$\begin{aligned} E[Y_i] &= W_q^{FCFS} \\ &= \frac{\lambda_i E[S^2]}{2(1-\rho_(i))}. \end{aligned} \tag{3.27}$$

Recall the work conserving principle

$$\begin{aligned} \sum_{j=1}^{i} \rho_j W_{jq}^{PR} &= \rho W_q^{FCFS} \\ &= \rho(i) \frac{\lambda(i) E[S^2(i)]}{2(1-\rho(i))}. \end{aligned} \tag{3.28}$$

Use (3.28) to find W_{1q}

$$\begin{aligned} \rho_1 W_{1q}^{PR} &= \rho(1) \frac{\lambda(1) E[S^2(1)]}{2(1-\rho(1))} \\ &= \frac{\rho_1 \lambda_1 E[S_1^2]}{2(1-\rho_1)}. \end{aligned} \tag{3.29}$$

and solve for W_{1q}^{PR}

$$W_{1q}^{PR} = \frac{\lambda_1 E[S_1^2]}{2(1-\rho_1)}. \tag{3.30}$$

Again, (3.28) is used to obtain W_{2q}^{PR}:

$$\begin{aligned}\rho_1 W_{1q}^{PR} + \rho_2 W_{2q}^{PR} &= \frac{\rho(2)\lambda(2)E[S^2(2)]}{2(1-\rho(2))} \\ &= \frac{\lambda(2)(\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2])}{2(\mu(2)-\lambda(2))} \\ &= \frac{(\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2])(\lambda_1 E[S_1] + \lambda_2 E[S_2])}{2[1-(\lambda_1 E[S_1] + \lambda_2 E[S_2])]}. \qquad (3.31)\end{aligned}$$

Solving for W_{2q}^{PR},

$$W_{2q}^{PR} = \frac{1}{\rho_2}\left\{\left[\frac{(\lambda_1+\lambda_2)^2 E[S]E[S^2]}{2(1-\rho)}\right] - \left[\frac{\lambda_1^2 E[S_1]E[S_1^2]}{2(1-\rho_1)}\right]\right\}. \qquad (3.32)$$

3.2 Results for M/G/1 Queue with Possible Rework

The following nomenclature applies to a multi-class M/G/1 priority queueing system with optional rework. In particular, arrivals are according to a Poisson process, and service times are general. After receiving a mandatory first service, a job of type $i, i = 1, 2, \ldots, k$, receives an optional second service with some non-zero probability, $\nu_i \in (0, 1]$.

S_i = random time to serve a job of type i; $i = 1, 2, ..., k$

T_i = random first service time for job of type i; $i = 1, 2, ..., k$

$G_i(\cdot)$ = cdf of first service time for job of type i; $i = 1, 2, ..., k$

Q_i = random second service time for job of type i; $i = 1, 2, ..., k$

$H_i(\cdot)$ = cdf of second service time for job of type i; $i = 1, 2, ..., k$

ν_i = probability job type i needs second service; $i = 1, 2, ..., k$.

Using a standard conditional expectation argument, the first and second moments of the service time for job type i are obtained in the following manner:

$$\begin{aligned}
S_i &= T_i + Q_i \\
E[S_i] &= E[T_i + Q_i | Q_i > 0]P\{Q_i > 0\} + E[T_i + Q_i | Q_i = 0]P\{Q_i = 0\} \\
&= E[T_i + Q_i | Q_i > 0]\nu_i + E[T_i + Q_i | Q_i = 0](1 - \nu_i) \\
&= \{E[T_i] + E[Q_i]\}\nu_i + E[T_i](1 - \nu_i) \\
&= E[T_i] + \nu_i E[Q_i] \qquad (3.33)
\end{aligned}$$

$$\begin{aligned}
S_i^2 &= (T_i + Q_i)^2 \\
E[S_i^2] &= E[(T_i + Q_i)^2 | Q_i > 0]P\{Q_i > 0\} + E[(T_i + Q_i)^2 | Q_i = 0]P\{Q_i = 0\} \\
&= E[(T_i + Q_i)^2 | Q_i > 0]\nu_i + E[(T_i + Q_i)^2 | Q_i = 0](1 - \nu_i) \\
&= E[T_i^2 + 2T_iQ_i + Q_i^2 | Q_i > 0]\nu_i + [T_i^2 + 2T_iQ_i + Q_i^2 | Q_i = 0](1 - \nu_i) \\
&= \{E[T_i^2] + 2E[T_i]E[Q_i] + E[Q_i^2]\}\nu_i + E[T_i^2](1 - \nu_i) \\
&= E[T_i^2] + \nu_i\{2E[T_i]E[Q_i] + E[Q_i^2]\}. \qquad (3.34)
\end{aligned}$$

For the aggregate measures, $E[S]$ and $E[S^2]$, equations (3.2) and (3.4) hold replacing the $E[S_i]$ and $E[S_i^2]$ with equations (3.33) and (3.34), respectively.

3.2.1 First-Come, First-Served Policy with Rework

For a single-station, multi-class M/G/1 priority queue with rework that follows a FCFS scheduling discipline, W_q, W_1, W_2, and L are found via the P-K formula

$$L = \rho + \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \qquad (3.35)$$

Dividing by λ yields

$$W^{FCFS} = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)}, \qquad (3.36)$$

which in turn yields the average waiting time in queue for an arbitrary job,

$$W_q^{FCFS} = \frac{\lambda E[S^2]}{2(1-\rho)} \tag{3.37}$$

where $E[S^2]$ is as defined in equation (3.34). Since the queue discipline is FCFS, the expected wait in queue is the same for jobs of class 1 and 2 and the total time in system for jobs 1 and 2 are, respectively,

$$W_1^{FCFS} = E[S_1] + W_q^{FCFS} \tag{3.38}$$

and

$$W_2^{FCFS} = E[S_2] + W_q^{FCFS}. \tag{3.39}$$

For completeness, the average number of jobs in the queue is

$$L_q = \frac{\lambda^2 E[S^2]}{2(1-\rho)}. \tag{3.40}$$

3.2.2 Non-Preemptive Priority Policy with Rework

The average waiting time in queue for a job of type i under the non-preemptive priority (NPP) policy is found using the same argument as in Section 3.1.2. Then, applying equation (3.18), it is seen that for a job of type 1, W_q is

$$E[W_{1q}^{NPP}] = \frac{\lambda E[S^2]}{2(1-\rho_1)}. \tag{3.41}$$

For $i = 2$,

$$\begin{aligned} W_{2q}^{NPP} &= \frac{\lambda E[S^2] + 2(\rho_1 W_{1q}^{NPP} + \rho_2 E[W_{2q}^{NPP}])}{2(1-\rho_1)} \\ &= \frac{\lambda E[S^2]}{2(1-\rho)(1-\rho_1)}. \end{aligned} \tag{3.42}$$

To obtain L_{1q} and L_{2q}, Little's Law would be applied giving

$$L_{1q} = \lambda_1 W_{1q}^{NPP} \tag{3.43}$$

and

$$L_{2q} = \lambda_2 W_{2q}^{NPP}. \tag{3.44}$$

3.2.3 Preemptive Resume Policy with Rework

Finally, the average waiting time in queue for a job of type 1 under the preemptive resume policy is found using the same argument as in Section 3.1.3. Since PR with rework is also work conserving, equations (3.30) and (3.32) can be used to find W_{1q} and W_{2q}. It follows that under PR policy with rework the following equations hold:

$$W_{1q}^{PR} = \frac{\lambda_1 E[S_1^2]}{2(1-\rho_1)} \tag{3.45}$$

and

$$W_{2q}^{PR} = \frac{1}{\rho_2}\left\{\left[\frac{(\lambda_1+\lambda_2)^2 E[S]E[S^2]}{2(1-\rho)}\right] - \left[\frac{\lambda_1^2 E[S_1]E[S_1^2]}{2(1-\rho_1)}\right]\right\}. \tag{3.46}$$

Again, to obtain L_{q1}, and L_{q2}, Little's Law would be applied yielding

$$L_{1q} = \lambda_1 W_{1q}^{PR} \tag{3.47}$$

and

$$L_{2q} = \lambda_2 W_{2q}^{PR}. \tag{3.48}$$

3.3 New Policy: Preemption Between First and Second Service Only

A new policy is proposed that allows preemptions to occur only between the mandatory first service and the optional second service. In other words, the queue is still ordered highest priority first, but the only time a type 1 job preempts a type

2 job is in between first and second service. It is hypothesized this policy costs less than the other three policies. This is a logical conjecture, since the set-up cost will be less than the PR policy, and the expected wait time, which contributes directly to the cost equations, for the higher class job type is less than the expected wait time for the FCFS and NPP policies. First, the average time in queue for a job of type 1 is derived, and then the average time in queue for a job of type 2 is derived.

3.3.1 Performance Measures for New Policy

The long run average time in queue for a job of type 1 is the average remaining service time of an arbitrary job plus the long run average time to serve all jobs of type 1 that are currently in the queue. To derive this mathematically, define:

$U_i =$ Random remaining service time of current job in service when a job of type i arrives, $i = 1, 2$

$Z =$ A random variable denoting type of job in service upon arrival.

$$Z = \begin{cases} 1 & \text{if type 1 is in service} \\ 2 & \text{if type 2 is in service} \end{cases}$$

$N =$ The service cycle in which the current job is undergoing.

$$N = \begin{cases} 1 & \text{if currently undergoing first service} \\ 2 & \text{if currently undergoing second service} \end{cases}$$

$X =$ Random number of jobs in system upon arrival in the long run.

The average wait in queue for a job of type 1 is

$$W_{1q}^{New} \quad = \quad E[U_1] + \rho_1 W_{1q} = \frac{E[U_1]}{(1-\rho_1)}. \tag{3.49}$$

Using equation (3.28), the average wait in queue for a job of type 2 is

$$W_{2q}^{New} \;=\; \frac{1}{\rho_2}\left[\frac{\lambda^2 E[S]E[S^2]}{2(1-\rho)} - \frac{\rho_1 E[U_1]}{(1-\rho_1)}\right]. \tag{3.50}$$

The expected remaining service time, $E[U_1]$ is found by conditioning on X:

$$E[U_1] \;=\; E[U_1|X>0]P(X>0) + E[U_1|X=0]P(X=0). \tag{3.51}$$

It is clear that $E[U_1|X=0]=0$, since there can be no remaining service time in an empty server. Now, $E[U_1|X>0]P(X>0)$ has to be derived. Again, this is accomplished through a conditional argument.

$$\begin{aligned} E[U_1|X>0] \;=\;& E[U_1|X>0, Z=1]P(Z=1|X>0)+ \\ & E[U_1|X>0, Z=2]P(Z=2|X>0). \end{aligned} \tag{3.52}$$

However, from Fakinos' [13] results of the remaining service time for the more general case of a GI/G/1 queue, the following equation holds for $E[U_1|X>0, Z=1]$.

$$E[U_1|X>0, Z=1] = \frac{E[S_1^2]}{2E[S_1]}. \tag{3.53}$$

To obtain $E[U_1|X>0, Z=2]$, another conditional expectation argument is applied. The remaining service of the current job in service depends on the job type currently being served and the service cycle the current job is undergoing. When a job of type 1 enters the system and sees a type 2 job in service, the type 2 job could be in first or second service. If the job of type 2 is in the first service, then the remaining service time of a job of type 2 when a job of type 1 arrives is the remaining service time of the first service of the job of type 2 currently in service. This is true since the job of type 1 preempts the job of type 2 before the job of type 2 enters second service. However, if an arriving type 1 job sees a type 2 job in second service, then the remaining service time is the expected remaining service time of the second service of a job of type 2. Also, the probability of a type 2 job being in first or second service is the proportion of time spent in first or second service, respectively, compared to the total amount of time a job of type 2 is either in service or waiting for the jobs of type 1 to be served that arrived during the first service of the type 2 job. The time also accounts for the time to serve all jobs of type 1 that arrive during the expected service time of the type 1 job that preempted the type 2 job.

The details needed to compute equation (3.52) are presented; again, a conditional expectation argument is applied.

$$\begin{aligned} E[U_1|X>0,Z=2] &= E[U_1|X>0,Z=2,N=1]P(N=1|X>0,Z=2)+ \\ &\quad E[U_1|X>0,Z=2,N=2]P(N=2|X>0,Z=2) \\ &= \left(\frac{E[T_2^2]}{2E[T_2]}\right)\left(\frac{E[T_2]}{E[S_2]+(\lambda_1 E[S_1])^2 E[T_2]}\right)+ \\ &\quad \left(\frac{E[Q_2^2]}{2E[Q_2]}\right)\left(\frac{\nu_2 E[Q_2]}{E[S_2]+(\lambda_1 E[S_1])^2 E[T_2]}\right) \\ &= \frac{E[T_2^2]+\nu_2 E[Q_2^2]}{2(E[S_2]+\rho_1^2 E[T_2])} \end{aligned} \tag{3.54}$$

Substituting equations (3.53) and (3.54) into equation (3.52), $E[U_1|X>0]$ can be found:

$$\begin{aligned} E[U_1|X>0] &= \frac{E[S_1^2]}{2E[S_1]}\frac{\rho_1}{\rho}+\left[\frac{E[T_2^2]+\nu_2 E[Q_2^2]}{2(E[S_2]+\rho_1^2 E[T_2])}\right]\frac{\rho_2}{\rho} \\ &= \frac{\lambda_1 E[S_1^2]}{2\rho}+\frac{\lambda_2 E[S_2](E[T_2^2]+\nu_2 E[Q_2^2])}{2\rho(E[S_2]+\rho_1^2 E[T_2])}. \end{aligned} \tag{3.55}$$

Finally, $E[U_1]$ is obtained by unconditioning on X. Using equation (3.55) and the fact that $E[U_1|X=0]=0$, the following result is obtained:

$$E[U_1] \quad = \quad \frac{\lambda_1 E[S_1^2]}{2} + \frac{\rho_2(E[T_2^2] + \nu_2 E[Q_2^2])}{2(E[S_2] + \rho_1^2 E[T_2])}. \tag{3.56}$$

Substituting equation (3.56) into equations (3.49) and (3.50) the final analytical results for the expected time in queue for a type 1 job and a type 2 job are obtained:

$$W_{1q}^{New} \quad = \quad \frac{1}{2(1-\rho_1)}\left[\lambda_1 E[S_1^2] + \frac{\rho_2(E[T_2^2] + \nu_2 E[Q_2^2])}{(E[S_2] + \rho_1^2 E[T_2])}\right], \tag{3.57}$$

$$W_{2q}^{New} \quad = \quad \frac{1}{2\rho_2}\left\{\frac{\lambda\rho E[S^2]}{(1-\rho)} - \left[\frac{\rho_1\lambda_1 E[S_1^2] + 2\rho_1\rho_2(E[T_2^2] + \nu_2 E[Q_2^2])}{(1-\rho_1)}\right]\right\}. \tag{3.58}$$

3.4 Long-Run Expected Cost Equations

To compare the various policies, the total long-run expected cost of each scheduling policy is used. The total cost consists of the holding cost for each job type plus the preemption costs. Clearly, there will only be preemption costs for policies that allow preemption. This section provides the appropriate cost equations for each of the four policies.

To derive the cost equations, define the following:

$\gamma_i =$ Holding costs for job type i

$\phi =$ Preemption cost during service

$\theta =$ Preemption cost between first and second service, where $\theta \ll \phi$

$\pi =$ Random variable denoting total number of preemptions per unit time

$D =$ Indicator variable indicating if the job currently in service requires second service. $D = 0$ if the job does not require second service and $D = 1$ if the job does require second service.

$C_i =$ Number of type i jobs processed per unit time

$A_i(t) =$ Total number of type i arrivals during $[0, t]$

$B =$ Random variable denoting the total number of times a job of type 2 is preempted between first and second service

$TC_{FCFS} =$ Long-run average total cost equation for FCFS discipline

$TC_{NPP} =$ Long-run average total cost equation for NPP discipline

$TC_{PR} =$ Long-run average total cost equation for PR discipline

$TC_{New} =$ Long-run average total cost equation for the new discipline

Any fixed costs for servicing a job are excluded since such costs will not alter the optimal solution.

3.4.1 Cost Equations for Existing Policies

The cost equation for FCFS includes only the holding cost, therefore, it is easily seen that, for the case $k = 2$,

$$TC_{FCFS} = \gamma_1 W_{1q}^{FCFS} + \gamma_2 W_{2q}^{FCFS}. \tag{3.59}$$

Similarly, the cost equation for the non-preemptive priority (NPP) discipline is straight-forward and only includes holding costs; therefore, for the $k = 2$ case

$$TC_{NPP} = \gamma_1 W_{1q}^{NPP} + \gamma_2 W_{2q}^{NPP}. \tag{3.60}$$

For the preemptive resume (PR) priority discipline, the cost equation includes the holding cost as well as the cost of preemption during service. The total cost equation is

$$TC_{PR} = \gamma_1 W_{1q}^{PR} + \gamma_2 W_{2q}^{PR} + \phi E[\pi], \tag{3.61}$$

where $E[\pi] = \lambda_1 E[S_2]$, which is the expected number of type 1 arrivals during one service cycle of a type 2 job.

3.4.2 Preemption Between First and Second Service

The total cost equation for the new policy also consists of both the holding cost and preemption cost. Again, the holding cost is straight-forward, however, the preemption cost needs to be derived. The expected preemption cost is the cost of preemption times the number of expected preemptions in the long-run multiplied by the number of type 2 jobs processed per unit time. Since θ, the cost for preemption, is known, $E[\pi]$, the expected preemption cost, is all that is needed. However, it is known that

$$E[\pi] = E[B]E[C_2], \tag{3.62}$$

where $E[B]$ is the long-run expected number of times a job of type 2 is preempted between first and second service and is derived by a simple conditional argument:

$$\begin{aligned} E[B] &= E[E[B|D]] \\ &= E[E[B|D=1]P(D=1) + E[B|D=0]P(D=0). \end{aligned} \tag{3.63}$$

It is known that $E[B|D = 0] = 0$, since there can be no preemptions if there is no second service. In addition, it is known that $E[B|D = 1]$ is the same as the expected number of type 1 arrivals during a type 2 job's first service cycle, given such a job requires second service. To obtain the expected number of type 1 arrivals during a type 2's first service cycle, $A_1(t)$, the total number of type 1 arrivals during $(0, t)$, is found by conditioning on T, where T is the first service time of job 2. Since only one job of type 1 needs to arrive, the complementary probability of $A_1(t)$ will be applied. The distribution of $A_1(t)$ is known to be Poisson:

$$\begin{aligned} P(A_1(t) = 0) &= \int_0^\infty P(A_1(t) = 0|T = t)dG_2(t) \\ &= \int_0^\infty e^{-\lambda_1 t} dG_2(t) \end{aligned} \tag{3.64}$$

and the complementary probability is

$$\begin{aligned} P(A_1(t) > 0) &= 1 - \int_0^\infty e^{-\lambda_1 t} dG_2(t) \\ &= 1 - \tilde{G}_2(\lambda_1) \end{aligned} \tag{3.65}$$

where $\tilde{G}_2(\lambda_1)$ is the Laplace-Stieltjes transform (LST) of $G_2(\cdot)$ evaluated at λ_1. Equation (3.65) can always be computed, as long as the LST is known. Now, $E[B]$ can be found:

$$E[B] = \nu_2(1 - \tilde{G}_2(\lambda_1)). \tag{3.66}$$

To find the preemptions per unit time, $E[C_2]$ is still needed. The expected type 2 jobs processed per unit time is equivalent to the expected arrivals of type 2 jobs per unit time, which is λ_2:

$$E[C_2] = \lambda_2. \tag{3.67}$$

Equation (3.62) can now be written, yielding the total cost equation for the new scheduling policy

$$TC_{New} = \gamma_1 W_{1q}^{New} + \gamma_2 W_{2q}^{New} + \theta\lambda_2\nu_2(1 - \tilde{G}_2(\lambda_1)). \tag{3.68}$$

A cursory observation of the four cost equations, (3.59), (3.60), (3.61), and (3.68), reveals the variable components are the policy-dependent mean queueing times and the fixed preemption costs. In the Chapter 4, numerical experiments are conducted to investigate which policy performs best over a wide range of problem parameters.

4. Numerical Results

For the four scheduling policies presented in Chapter 3, the analytical cost equations are compared across the policies, and the steady-state expected times in queue are compared to an ARENA simulation.

4.1 Benchmark Analytical Results

Before considering explicit comparisons of the four scheduling policies, an ARENA simulation was used to benchmark the analytical results. Ten scenarios were constructed varying the problem parameters.

The values of the varying parameters were chosen at random. In all cases the stability condition is satisfied and every entity has a non-zero probability of rework. The first and second service times were drawn from the exponential distribution. Thus, if an entity needed rework, the total service time was the sum of two exponential random variables. The parameter variations for the ten test scenarios are summarized in Table 4.1. These parameters were used in simulation experiments of all four policies. The warm-up period was 2,000 minutes, the replication length was 100,000 minutes, and 25 replications were performed. The warm-up period was estimated using ARENA's output analyzer graphics tool. In order to estimate when the system stabilized, the work in process (WIP) of both types of jobs was plotted over time. The run length was chosen to reduce the half-width within a replication for the long-run expected time in queue, and the number of replications was chosen to reduce the half-width across replications.

Table 4.1 Summary of problem parameters for 10 scenarios.

Case	λ_1	λ_2	$E[T_1]$	$E[T_2]$	$E[Q_1]$	$E[Q_2]$	ν_1	ν_2
1	1.0001	1.1111	0.1000	0.1500	0.6428	0.3000	0.7000	0.4000
2	1.2000	0.7000	0.3500	0.3000	0.4000	0.2000	0.6000	0.3000
3	0.4000	0.8500	0.1500	0.1000	1.5833	0.2157	0.3500	0.7000
4	0.2000	0.5850	0.2000	0.5000	0.7000	0.5650	0.5000	0.6500
5	0.6000	0.2000	0.9650	0.2000	0.5000	0.7750	0.7318	0.1615
6	0.3000	0.6000	0.7000	0.9000	0.7350	0.4750	0.2500	0.3000
7	0.5440	0.7000	0.3500	0.3000	0.4000	0.2000	0.6000	0.3000
8	0.1670	0.2000	1.2000	2.0000	0.8000	0.7000	0.6500	0.4500
9	0.2337	0.3371	0.3000	0.4750	1.0599	0.3937	0.5500	0.3209
10	0.7148	1.7162	0.1500	0.1000	0.4000	0.3580	0.6000	0.4000

The mean waiting time in queue for each job type was computed and compared to the analytical mean waiting times in queue for each job type. Tables 4.2 through 4.5 provide a summary of the results.

Table 4.2 Analytical versus simulated results for FCFS policy.

	W_{1q}			W_{2q}		
Case	Analytical	Simulated	Half-Width	Analytical	Simulated	Half-Width
1	2.8616	2.8600	0.0342	2.8616	2.8615	0.0344
2	11.1750	11.1847	0.4502	11.1750	11.1742	0.4113
3	0.8757	0.8760	0.0084	0.8757	0.8760	0.0092
4	1.1657	1.1651	0.0138	1.1657	1.1664	0.0128
5	6.6904	6.7320	0.1249	6.6904	6.7360	0.1232
6	7.5855	7.5706	0.2339	7.5855	7.5711	0.2344
7	0.5821	0.5830	0.0055	0.5821	0.5838	0.0047
8	5.5424	5.5725	0.1066	5.5424	5.5748	0.1173
9	0.5404	0.5428	0.0062	0.5404	0.5437	0.0054
10	0.7904	0.7937	0.0065	0.7904	0.7935	0.0061

As can be seen in Table 4.2, the simulated mean time in queue for the FCFS policy is within 0.62 percent of the true long-run mean time in queue.

As can be seen in Table 4.3 the analytical long-run expected time in queue for jobs of type 1 and type 2 are within the half-widths of the simulated long-run expected times in queue, with the maximum error being 1.45 percent.

Table 4.3 Analytical versus simulated results for NPP policy.

	W_{1q}			W_{2q}		
Case	Analytical	Simulated	Half-Width	Analytical	Simulated	Half-Width
1	0.9539	0.9560	0.0063	6.3592	6.4645	0.1053
2	1.5308	1.5335	0.0127	38.2705	38.8350	1.9220
3	0.6156	0.6205	0.0076	1.2191	1.2227	0.0169
4	0.5012	0.5003	0.0030	1.3097	1.3109	0.0149
5	4.5306	4.5422	0.0631	33.2088	33.5580	1.1360
6	1.1290	1.1286	0.0083	10.3222	10.3069	0.3472
7	0.3660	0.3657	0.0025	0.8572	0.8594	0.0079
8	1.9421	1.9453	0.0158	7.7759	7.7952	0.1524
9	0.4024	0.4036	0.0040	0.6809	0.6838	0.0074
10	0.3330	0.3330	0.0016	1.0959	1.0964	0.0072

Table 4.4 displays the simulation results and the analytical results for steady-state expected time in queue for jobs of type 1 and type 2 under the PR policy. In the PR policy simulation, the maximum relative error is 2.13 percent. Table 4.5 displays the simulated and analytical expected times in queue for the new policy. The relative error for this simulation was 5.05 percent. The difference in the expected times in queue can be attributed to the randomness of the simulation. The half-widths in Tables 4.2 through 4.5 are for a 95 percent confidence interval.

4.2 Determining the "Best" Policy

A numerical experiment was performed to determine which policy yields the lowest cost most often. To accomplish this, the long-run total expected cost was calculated and compared for 288 different scenarios. The comparison of the "best" policy was based on frequency of "wins" since this measure captures the number of times a particular policy yielded the lowest cost over a variety of situations. The following parameters were varied: ρ_1, ρ_2, γ_1, γ_2, ϕ, θ, ν_1, and ν_2. *MATLAB*® was used to compute the total long-run expected cost for the 288 scenarios and

Table 4.4 Analytical versus simulated results for PR policy.

	W_{1q}			W_{2q}		
Case	Analytical	Simulated	Half-Width	Analytical	Simulated	Half-Width
1	0.7650	0.7674	0.0057	6.7055	6.7951	0.1066
2	1.2432	1.2469	0.0121	39.0788	39.7107	1.9270
3	0.5474	0.5518	0.0075	1.3092	1.3217	0.0175
4	0.0798	0.0792	0.0012	1.4011	1.4179	0.0152
5	4.3697	4.3836	0.0614	35.1841	34.8366	1.1460
6	0.3077	0.3073	0.0043	10.6703	10.6827	0.3487
7	0.2423	0.2424	0.0021	1.0147	1.0297	0.0085
8	0.5811	0.5795	0.0089	8.6203	8.7268	0.1555
9	0.2599	0.2624	0.0035	0.8259	0.8394	0.0083
10	0.1531	0.1532	0.0011	1.2160	1.1906	0.0074

to calculate which policy resulted in the lowest cost most often. Equations (3.59), (3.60), (3.61), and (3.68) of Chapter 3 were used for this purpose.

The expected first service times and the arrival rates for each job type were kept constant throughout the 288 scenarios. The expected second service times were calculated using the fixed first service times, ρ_i and ν_i, respectively. The following formula was used to compute the mean second service time,

$$E[Q_i] = \frac{\rho_i - \lambda_i E[T_i]}{\lambda_i \nu_i}, \quad i = 1, 2. \tag{4.1}$$

where $E[T_1]$ =0.1, $E[T_2]$ =0.15, λ_1 =0.5, and $E[T_1]$ =0.65. The different values for the varying parameters were chosen to cover a wide range of situations. In all cases, the cost to preempt during service (θ) is greater than the cost to preempt between services (ϕ), since it is assumed to be cheaper to preempt between services than it is to preempt during service. Also, the holding cost for jobs of type 1 (γ_1) is greater than the holding cost for jobs of type 2 (γ_2), since a higher priority job is assumed to have a more expensive holding cost. Realistically, this makes sense because a higher priority is often given to the job with the higher holding cost. Three sets of

Table 4.5 Analytical versus simulated results for the new policy.

	W_{1q}			W_{2q}		
Case	Analytical	Simulated	Half-Width	Analytical	Simulated	Half-Width
1	0.8887	0.9120	0.0062	6.4788	6.5374	0.1056
2	1.4156	1.4909	0.0126	38.5942	39.0520	1.9230
3	0.5962	0.6024	0.0076	1.2447	1.2505	0.0171
4	0.3784	0.3799	0.0023	1.3364	1.3509	0.0150
5	4.4674	4.5181	0.0626	33.9844	33.6840	1.1390
6	0.9833	1.0234	0.0075	10.3840	10.4044	0.3478
7	0.3392	0.3472	0.0024	0.8914	0.9021	0.0082
8	1.6865	1.7689	0.0155	7.9345	8.1563	0.1536
9	0.3731	0.3781	0.0038	0.7107	0.7226	0.0076
10	0.2944	0.2989	0.0015	1.1217	1.1120	0.0072

parameters were varied for the experiment: the utilization factors, the probability of rework, and the costs.

The values chosen for the parameters were such that a variety of traffic intensities, probabilities of rework, and costs were examined. For the overall traffic intensity, a high, medium and low value was chosen. The traffic intensity was then separated between the class types to observe the system with balanced and unbalanced traffic intensities across classes. The probabilities of rework were also chosen to observe values that were high, medium, and low. Finally, the costs were chosen to observe the systems with high and low holding costs, as well as high and low preemption costs. Tables 4.6 through 4.8 display the parameter set variations. A numerical experiment was chosen over a full-factorial designed experiment because a full-factorial designed experiment is used to predict a response variable and observe which parameters impact the response variable. However, this thesis compares expected total cost across four scheduling policies, therefore, the parameters that impact the expected total cost are the parameters that comprise the long-run expected times in queue. For this reason, a numerical experiment is best suited to accomplish the goals of this thesis. Table 4.6 displays the various parameters used for the class and overall traffic intensities. As discussed previously, there is a low,

Table 4.6 Class and overall traffic intensities.

ρ_1	ρ_2	ρ
0.250	0.250	0.500
0.400	0.100	0.500
0.100	0.400	0.500
0.375	0.375	0.750
0.600	0.150	0.750
0.150	0.600	0.750
0.475	0.475	0.950
0.700	0.250	0.950
0.250	0.700	0.950

medium, and high value for the overall traffic intensity. The rework probabilities are displayed in Table 4.7 and were chosen at levels considered high rework probabilities, low rework probabilities, and imbalanced rework probabilities. The cost parameters are displayed in Table 4.8. The holding cost for a job of type 1 is varied over high and low, while the holding cost for a job of type 2 was varied over high, medium and low. The preemption cost during service was also varied over high and low, while the preemption cost between service was varied over high, medium and low.

Table 4.7 Class-specific probability of rework.

ν_1	ν_2
0.9	0.9
0.9	0.3
0.3	0.9
0.2	0.2

Table 4.8 Summary of cost parameters.

γ_1	γ_2	ϕ	θ
50	5	10	1
50	5	10	5
50	25	10	1
50	25	10	5
10	1	50	5
10	5	50	5
10	1	50	25
10	5	50	25

After values were chosen for the three parameter sets, all combinations of the parameter sets were used as inputs to a $MATLAB^{®}$ computer program. The $MATLAB^{®}$ program computed the long-run average waiting time in queue and the total long-run average cost for each case, then compared the costs across the four scheduling policies. Each time a particular policy had the lowest cost, it was declared the “winner”. The results of the numerical experiment are summarized in Table 4.9.

Table 4.9 Results of the policy comparison experiment.

Scheduling Policy	Number of times won
First-come, first-served (FCFS)	48
Non-preemptive priority (NPP)	43
Preemptive resume (PR)	116
New policy (New)	81

As seen in Table 4.9, the PR policy yields the lowest cost most often, while the NPP policy yields the lowest cost least often. After reviewing the results, some conjectures were made. The first conjecture is that the FCFS policy always won when the holding cost of a type 1 job (γ_1) was twice the holding cost of a type 2 job (γ_2) and the traffic intensity for a type 1 job was strictly greater than the traffic intensity for a type 2 job. This indicates that when the server is more busy with type 1 jobs than type 2 jobs, and it is at least twice as expensive to hold a type 1 job than a type 2 job, then FCFS is the scheduling policy with the lowest expected cost. It was also found that, in general, when the holding cost for type 1 is low, and the cost to preempt during service is high, the new policy prevailed. This conjecture makes sense in the real-world since there is a trade-off between a high holding cost for jobs of type 1 and high costs due to preemption. Conversely, if the cost to preempt during service is low, or the probability of rework is low, then the PR policy yielded the lowest cost. Another conjecture is that when the overall traffic intensity is high, either the PR policy or the new policy yields the lowest cost most often. This conjecture suggests that the new policy or the PR policy is worthwhile to investigate when the system traffic intensity is high. Additionally, it appears there is interaction between several factors that impact the mean waiting times. For example, when the traffic intensity is equal for the two classes, no policy dominates the others.

4.3 Summary of Numerical Experiments

A numerical experiment was performed to determine the scheduling policy that yields the lowest cost most often. There were 288 test cases with varying parameters. The parameters covered a variety of scenarios, however in all cases it was assumed that it is less expensive to preempt between services than to preempt during service, and the holding cost for jobs of type 1 was always greater than the holding cost for jobs of type 2. After the costs were compared over all four scheduling policies, it was found that the preemptive resume (PR) policy resulted in the lowest cost most often. However, the new policy resulted in the second lowest cost most often.

The results of this experiment indicate that the PR policy minimizes cost for a two-class M/G/1 priority queue with optional rework more often than the other three scheduling policies. However, the new policy produces the lowest cost almost as often as the PR policy. The results of this experiment suggest that, relative to the FCFS and NPP policies, the PR and the new policy perform better. To minimize long-run expected total costs, it is worthwhile to assign priorities and employ a preemption scheme when scheduling multi-class M/G/1 queues. In reference to the avionics backshop described in Chapter 1, knowing the best policy to employ in scheduling line replaceable units (LRUs), could potentially save the USAF significant dollar amounts. The chosen parameter values directly impact which policy yields the lowest cost. Some conjectures were made concerning the impact of the different factors. These could be further investigated through a sensitivity analysis. The next chapter provides conclusions of this thesis and some suggestions for future work in this area.

5. Conclusions and Future Research

The objective of this thesis was to derive performance measures for a two-class M/G/1 priority queue with optional rework and to use these results to compare the performance of four scheduling policies. This analysis provides insight into many real-world systems in manufacturing, telecommunications, and transportation. Since the cost equations depend explicitly on the time spent in queue, the long-run expected time in queue was derived for all policies. Specifically, the total long-run expected cost for first come, first served (FCFS), non-preemptive priority (NPP), preemptive resume (PR) and a new policy that only allows preemptions between first and second service were compared. All of the relevant performance measures were found using a mean value analysis (MVA) approach, the cost equations were derived, and a numerical experiment was performed to heuristically determine the policy that most often yields the lowest cost. The analytical results were initially validated via Monte-Carlo simulation. This chapter summarizes what has been learned in this thesis and makes recommendations for future work in this area.

The results presented in Chapter 3 provide the mean waiting time in queue for the four different scheduling policies. An MVA approach was taken to derive the first and second moments of the service time distribution of an arbitrary arrival of a particular class, and for an arbitrary arrival of an arbitrary class. After the overall steady-state expected service times were derived, the long-run expected queueing time for an arbitrary customer was developed for each of the four scheduling policies, FCFS, NPP, PR and a new policy that only allows preemption between the first and second service cycles. The new policy allowed preemption strictly between first and second service. For the new policy, as well as the previous three policies, the long-run expected costs were also computed. Holding costs and (where appropriate) preemption costs were considered for each policy.

Chapter 4 provided numerical results for the cost comparisons. Several parameters were varied over a range of values to test the different policies. Specifically, three sets of parameters were varied: the utilization factors, the probabilities of rework, and the holding and preemption costs. In 288 test cases, the PR policy performed the best most often. The new policy performed almost as well, while the NPP policy performed the worst. Some conjectures were made concerning these results. It was found that, when the holding cost for a type 1 job was at least twice the holding cost of type 2 job and the traffic intensity of type 1 jobs was strictly greater than the traffic intensity of type 2 jobs, FCFS had the lowest cost. For the PR and the new policy cases, no definitive conjectures could be made, although some were hypothesized. These conjectures included the new policy generally winning when the holding cost for type 1 was low, and the cost to preempt during service was high. It was also found that, in general, when overall traffic intensity was high, the new policy or the PR policy prevailed. It was also hypothesized that when the cost to preempt during service was low, or the probabilities of rework were low, the PR policy yielded the lowest cost. These findings suggests that to minimize cost, it is beneficial to use a preemptive policy when scheduling multi-class M/G/1 priority queueing systems. Another conjecture was that there exists an interaction between the factors involved in the mean queueing times.

The contributions of this thesis are useful in a variety of settings, including transportation, communications, or manufacturing systems. The research could also apply to personnel deployment lines or a maintenance schedule, as discussed in Chapter 1. The multi-class M/G/1 queue with optional second service can be studied further. For example, sensitivity analysis can be performed on each of the parameters. The values and ranges different parameters may assume without changing the outcome for the lowest cost policy could be explored to evaluate the sensitivity of the cost equations. For example, if the PR policy prevails over the new policy in a particular case, the question of how much the costs can vary without changing the outcome could be explored. Additionally, finding the optimal first or second mean service rates would also be advantageous. If optimal rates were known, design engineers could plan real-world systems with this knowledge, thereby minimizing long-run costs.

Bibliography

1. Artalejo, J. R. (1998). Some results on the M/G/1 queue with N-policy. *Asia-Pacific Journal of Operational Research*, **15** 147-157.

2. Baskett, F., K. M. Chandy, R. R. Muntz, and F. G. Palacios (1975). Open, closed and mixed networks of queues with different classes of cutsomers. *Journal of the Association for Computing Machinery*, **22**, (2), 248-260.

3. Ball, J. A., M.V. Day, and P. Kachroo (1999). Robust feedback control of a single server queueing system. *Mathematics of Control, Signals and Systems*, **12**, 307-345.

4. Bard, J. F., K. Srinivasan, and D. Tirupati (1999). An optimization approach to capacity expansion in semiconductor manufacturing facilities. *Int. Journal of Production Research*, **15**, 3359-3382.

5. Bryant, R. M., A.E. Krzesinske, and P. Teunissen (1983). The MVA pre-empt priority approximation. *Proceedings ACM SIGMETICS Conference on Measurement and Modeling of Computer Systems*, 12-27.

6. Bryant, R. M. and A. Krzesinski (1984). The MVA approximation. *ACM Transactions on Computer Systems*, **2**, no. 4, 335-359.

7. Cox, D. R. (1955). The analysis of non-Markovian stochasitc processes by the inclusion of supplementary variables. *Proc. Camb. Phil. Soc.*, **51**, 433-441.

8. Dai, J. G. (1995). On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid models. *Ann. Applied Probability*, **5**, 49-77.

9. Day, M. V., J. Hall, J. Menendez, D. Potter, and I. Rothstein, (2002). Robust optimal service analysis of single-server re-entrant queues. *Computational Optimization and Applications*, **22**, 261-302.

10. Diaz-Rivera, I. and D. Armburster and T. Taylor (2000). Periodic orbits in a class of re-entrant manufacturing systems. *Mathematics of Operations Research*, **25**, (4), 708-725.

11. Doshi, B. T., (1986). Queueing systems with vacations - a survey. *Queueing Systems*, **1**, 29-66.

12. Eager, D. L. and J. N. Lipscomb, (1988). The AMVA priority approximation. *Performance Evaluation*, **8**, (3), 173-193.

13. Fakinos D. (1982). The expected remaining service time in a single server queue. *Operations Research*, **30**, (5), 1014-1018.

14. Green, L. (1982). A limit theorem on subintervals of interrenewal times. *Operations Research*, **30**, 210-216.

15. Groenevelt, R., G. Koole, and P. Nain (2002). On the bias vector of a two-class preemptive priority queue. *Mathematical Methods of Operations Research*, **55** 107-120.

16. Grosfeld-Nir, A. and Y. Gerchak (2002). Multistage production to order with rework capability. *Management Science*, **48**, 652-664.

17. Gross, D. and C. Harris (1985). *Fundamentals of Queueing Theory.* John Wiley and Sons, Inc., New York.

18. Jaiswal, N.K. (1968). *Priority Queues.* Academic Press.

19. Kendall, D. G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society, Ser. B*, **13**, 151-185.

20. Kelton, W. D., R. P. Sadowski, and D. A. Sadowski (2002). *Simulation with ARENA.* McGraw -Hill Companies, Inc.

21. Kleinrock, L. (1975). *Queueing Systems, Volume I: Theory.* John Wiley and Sons, Inc., New York.

22. Kleinrock, L. (1976). *Queueing Systems, Volume II: Computer Applications.* John Wiley and Sons, Inc.

23. Kumar, P. R. (1993). Re-entrant lines. *Queueing Systems: Theory and Applications: Special Issue on Queueing Networks*, **13**, 87-110.

24. Kumar, P. R. and D. Arivudainambi (2001). An M/G/1/1 feedback queue with regular and optional services. *Information and Management Sciences*, **12**, 67-73.

25. Lin, C., J. Yeh, and J. Ding (1998). Design of random inspection rate for a flexible assembly system: A heuristic genetic algorithm approach. *Microelectronics and Reliability*, **38**, (4), 545-551.

26. Little, J. D. C. (1961). A proof of the queueing formula: $L = \lambda W$. *Operations Research*, **9**, 383-387.

27. Madan, K. (2000). An M/G/1 queue with second optional service. *Queueing Systems*, **34**, 37-46.

28. Narahari, Y. and L. M. Khan (1996). Performance analysis of scheduling policies in re-entrant manufacturing system. *Computers and Operations Research*, **23**, (1), 37-51.

29. Park Y., S. Kim, and C. H. Jun (2002). Mean value analysis of re-entrant line with batch machines and mult-class jobs. *Computers and Operations Research*, **29**, 1009-1024.

30. Park Y., S. Kim, and C. H. Jun (2002). Performance evaluation of re-entrant lines with multi-class jobs and multi-server workstations. *Production Planning and Control*, **13**, 56-65.

31. Prabhu, N. U. (1975). *Queues and Inventories: A study of Their basic Stochastic Processes*. John Wiley and Sons, Inc., New York.

32. Saaty, T. (1961). *Elements of Queueing Theory*. McGraw-Hill, New York.

33. Stidham, Jr., S. (2002). Analysis, design and control of queueuing systems. *Operations Research*, Special 50th Anniversary Issue.

34. Zargar, A. M. (1995). Effect of rework strategies on cycle time. *Computers and Industrial Engineering*, **29**, (4), 239-243.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)* 25-03-2003	**2. REPORT TYPE** **Master's Thesis**	**3. DATES COVERED** *(From – To)* Jun 2002 – Mar 2003

4. TITLE AND SUBTITLE

ANALYSIS OF SCHEDULING POLICIES FOR A M/G/1 QUEUE WITH REWORK

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hendrixson, Jennifer K., 1Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/03-09

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This thesis analyzes a multi-class M/G/1 priority queueing system in which distinct job types require one service cycle and, with non-zero probability, require a second service cycle. The main objective is to find a new heuristic scheduling policy that minimizes the long-run expected holding and preemption costs. Arrival rates, service rates, and the probability of undertaking second service are all class specific. A mean value analysis (MVA) approach was employed to derive the long-run mean time in queue for each job type under each policy, thereby providing the appropriate cost equations. Numerical experiments suggest that the preemptive resume scheduling policy yields the lowest cost most frequently.

15. SUBJECT TERMS

M/G/1 Queue, Rework, Scheduling

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Jeffrey P. Kharoufeh, Ph.D. (ENS)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4336; e-mail: Jeffrey.Kharoufeh@afit.edu
U	U	U	UU	55	